Oplossingen voor het beheren van metadata

Metadata in de praktijk

Rick Mutsaers

Door groei van de vraag naar informatie en verdergaande integratie tussen systemen, worden metadata en het goed beheren van metadata ook in toenemende mate belangrijker. In dit artikel worden de verschillende vormen van metadata belicht waar we als gebruiker en ontwikkelaar/beheerder in de praktijk mee te maken hebben, en de manier waarop deze metadata het best kunnen worden vastgelegd en beheerd.

De gebruiker verdrinkt in de vele definities van business-data die er binnen een organisatie in de verschillende informatiesystemen zijn opgeslagen. Bovendien zijn ontwikkelaars en beheerders continu op zoek naar de beste bron voor data. Data worden immers vaak in meerdere systemen opgeslagen (zelfs bij gebruik van ERP-systemen) en het is altijd de vraag in welk systeem de betreffende data kwalitatief het beste zijn (het best gevuld c.q. het meest actueel). Metadata vormen daarvoor de sleutel.

Integratie

Elk systeem op zich beschikt wel over voldoende metadata, nodig voor de eigen correcte werking, alleen is een overall beeld over alle systemen heen vaak lastig samen te stellen. Ook is het nagenoeg onmogelijk om efficiënt door al deze bronnen van metadata te zoeken naar waar bijvoorbeeld een bepaald data-element gebruikt wordt. Goed beheer en integratie van de metadata uit al deze systemen wordt daarom belangrijker.

Uitwisseling

SuperGlue is gebaseerd op het Common Warehouse Metamodel en de Meta Object Facility, een set van afspraken over de integratie en uitwisseling van metadata tussen diverse tool-omgevingen. Beide zijn door de Object Management Group in het leven geroepen om de uitwisseling van metadata tussen diverse producten te verbeteren. Het definieert op welke manieren en met welke datatypes metadata uitgewisseld zouden moeten worden om deze gemakkelijk in diverse tools te kunnen inlezen. Daardoor kunnen bijvoorbeeld datamodellen die gemaakt zijn met een tool als ERwin of PowerDesigner, eenvoudig ingelezen worden in ETL tools en reporting tools.

Instrumenten waarmee metadata opgeslagen, beheerd en geïntegreerd kunnen worden, komen inmiddels langzamerhand op de markt. Dit artikel gaat specifiek wat dieper in op de mogelijkheden die dit soort tools bieden aan IT-afdelingen.
Er zijn ook tools op de markt die zich alleen richten op het samenstellen van een business metadata repository (zoals bijvoorbeeld Britcom's Infolibrarian), maar deze worden voornamelijk gebruikt door eindgebruikers die definities en eigenaars willen kunnen zoeken van gegevens binnen de organisatie of de actualiteit van de gegevens in een rapport.

Uit een onderzoek van Gavilan Research blijkt dat 83 procent van de ondervraagde bedrijven metadata tools wil aanschaffen om impact-analyses uit te voeren en ruim tweederde van de respondenten wil data lineage kunnen doen en transformatie-regels, mappings en processen documenteren. Dit zijn voornamelijk IT-gerelateerde issues. Vandaar dat dit artikel met name ingaat op metadata tools die primair gericht zijn op IT-afdelingen. Informatica's SuperGlue is daar een mooi voorbeeld van; aan de hand van wat voorbeelden van het gebruik van SuperGlue worden de verschillende aspecten van metadata-beheer belicht.

Soorten metadata

Er bestaan verschillende soorten metadata. Volgens de definitie zijn metadata gegevens over de gegevens, met andere woorden ze zeggen iets óver de betreffende gegevens. Dat 'iets' kan bijvoorbeeld een business-definitie van winst zijn of de betekenis van de term 'klant'; maar ook uit welke onderliggende data het gegeven 'winst' is opgebouwd; of wanneer we spreken van een klant in plaats van een prospect; of wie verantwoordelijk is voor het vastleggen en de kwaliteit van de klantgegevens. Dit zijn allemaal beschrijvende, business-gerichte metadata.

Daarnaast bestaan er de meer technische metadata, zoals de definitie van een database-tabel met zijn kolommen en de datatypes van deze kolommen; of de actualiteit van de gegevens in die tabel. Vragen die daarbij spelen: welke waarden mag een kolom bevatten (domeingrenzen); hoe ziet een ETL-proces eruit, welke bronnen worden gebruikt; hoe worden data getransformeerd en in welke doeltabel komen deze data daarna terecht.
Als laatste soort bestaan er dan nog procesmatige of operationele metadata die iets zeggen over de uitvoering van gegevensverwerkende processen, zoals bij ETL tools. Daarbij worden zaken vastgelegd als wanneer een ETL-proces gestart is en hoelang het duurde, maar ook of er fouten zijn opgetreden en hoeveel records van bron naar doelsysteem zijn verwerkt. Met dit soort metadata kan ook de actualiteit en kwaliteit van gegevens worden weergegeven.

Beheer van metadata

Al deze metadata zitten over het algemeen in een repository van het bijbehorende tool. Dus metadata over fysieke structuur van gegevenselementen zit in de repository van het database management-systeem (ook wel data dictionary genoemd). Metadata over de rapportelementen in een reporting tool zitten in de catalog of universe of hoe het ook in het betreffende reporting tool heet. Bron- en doeldefinities en transformatieregels van een transformatieproces zitten in de repository van een ETL tool, naast de procesgegevens van de uitvoering van de transformatieprocessen.
Kortom, de metadata zíjn er wel, ze zijn alleen slechts geïsoleerd beschikbaar. In de data dictionary van een database zijn geen metadata te vinden uit het reporting tool waar de databasegegevens op voorkomen. En in een reporting tool zijn geen metadata te vinden over de transformatieprocessen waarmee gegevens uit bronsystemen in een datawarehouse worden geladen.

> Om problemen te verhelpen komen er vaker gespecialiseerde metadata tools op de markt

Door deze versnippering is het voor een gebruiker lastig te bepalen of de gegevens die hij op zijn wekelijkse omzetrapport terugvindt, wel actueel en juist zijn. De daarvoor benodigde procesinformatie zit immers niet in de metadata van het reporting tool, maar in de metadata van het gebruikte ETL tool. Voor de ontwikkelaar is het lastig te bepalen uit welke brontabel een bepaald business gegeven moet worden gehaald. Voor de wijzigingsbeheerder is het een schier onmogelijke taak de impact van een wijziging van een database-tabel in een bronsysteem te bepalen die het op de rest van de informatie-infrastructuur heeft.
Om deze problemen te verhelpen komen er vaker gespecialiseerde metadata tools op de markt, waarmee de individuele repository's van tools gelezen en geïntegreerd kunnen worden tot één centrale metadata-repository. Deze tools lezen doorgaans metadata uit de repository's van diverse soorten database management-systemen, reporting tools en ETL tools en plaatsen die in een centrale metadata-repository. Daarbij wordt een aantal analyse- en rapportagemogelijkheden geboden (meestal in de vorm van wat reports) om deze centrale metadata te ontsluiten. Afbeelding 1 toont een rapportagescherm van SuperGlue, één van de metadata tools die op dit moment op de markt zijn.

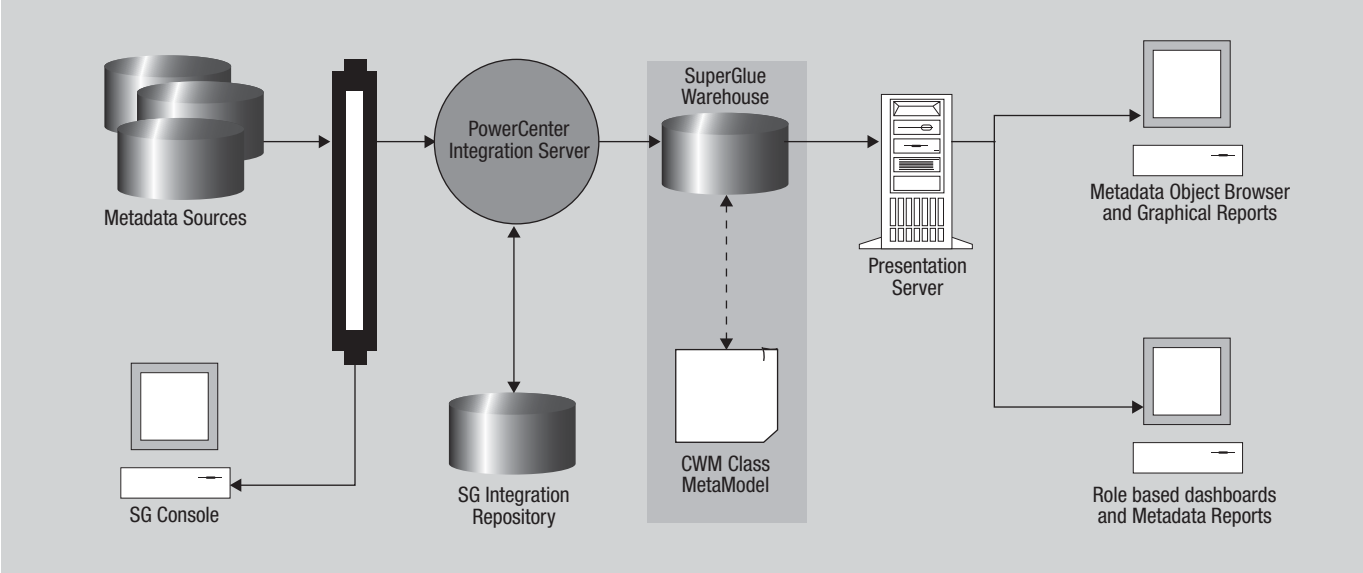
Afbeelding 1: Voorbeeld van operationele metadata-rapportage in SuperGlue.

Daardoor kan de eindgebruiker zien uit welk bronsysteem zijn rapportgegevens komen en hoe actueel de gegevens zijn.
De ontwikkelaar kan voor het samenstellen van een nieuw rapport op basis van een business-definitie snel opzoeken in welk bronsysteem en daarbinnen in welke tabel en kolom, deze gegevens zijn opgeslagen en de wijzigingsbeheerder kan de impact bepalen van een wijziging van één van de systemen.
Uiteraard zijn er legio mogelijkheden en is deze lijst van mogelijkheden slechts een greep.

Werking metadata-tool SuperGlue

De meeste metadata-tools zijn opgebouwd uit de volgende basiscomponenten:

- Een metadata repository database waarin de geconsolideerde metadata worden opgeslagen;
- Een mechanisme om metadata uit diverse soorten tool repository's te lezen en in de metadata repository te stoppen;



Afbeelding 2 : Architectuur van SuperGlue.

- Een reporting- en analyse-omgeving waarmee de geconsolideerde metadata ontsloten kunnen worden.

In het voorbeeld van SuperGlue betreft het dan de volgende componenten: SuperGlue Warehouse; Xconnects i.s.m. PowerCenter; PowerAnalyzer.

SuperGlue warehouse is de plek waar in SuperGlue alle metadata worden opgeslagen: de kern van het systeem. Dit warehouse is gebaseerd op het Common Warehouse Model (CWM) en de Meta Object Facility (MOF) van de Object Management Group (OMG) (zie kader), een standaard voor het generiek opslaan en uitwisselen van metadata tussen diverse DBMS'en, ETL tools en reporting tools.

> Naast de uit een tool afkomstige metadata kan SuperGlue warehouse ook uitgebreid worden met 'eigen' metadata

In SuperGlue Warehouse worden de ingelezen metadata opgeslagen met behulp van de termen zoals ze ook in het betreffende tool zijn opgeslagen. Zo wordt bij de metadata uit een BusinessObjects repository gesproken over universes en measures en bij metadata uit een PowerCenter repository wordt gesproken over mappings en workflows; herkenbare termen dus.
In de basis zijn het echter gestandaardiseerde metadata-items. Een source-definitie in PowerCenter is een table in de Oracle data dictionary en een class in een BusinessObjects repository. Dit zijn drie verschillende termen dus, die, wanneer in SuperGlue een data lineage uitgevoerd wordt, aan elkaar gerelateerd kunnen worden dankzij hun onderliggende betekenis als tabeldefinitie. Naast met de uit een tool afkomstige metadata kan het SuperGlue warehouse ook uitgebreid worden met 'eigen' metadata, bijvoorbeeld welke afdeling er verantwoordelijk is voor een mapping of een database-tabel. Dit zijn doorgaans geen metadata die in een tool repository worden vastgelegd.

Xconnects en PowerCenter

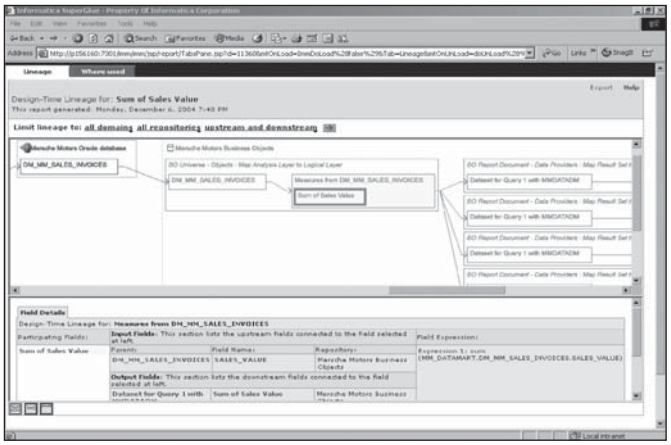
Met behulp van XConnects wordt de repository van de diverse ondersteunde tools gelezen. Een XConnect is eenvoudig gesteld een set van query's om metadata uit een tool repository te halen. Vervolgens wordt PowerCenter gebruikt om de metadata met behulp van de XConnects uit de tool repository's te lezen en te schrijven in het SuperGlue Warehouse.
In de laatste versie van SuperGlue (2.1) worden de volgende XConnects ondersteund:

- Databases (Oracle 8.1.7, 9i, 10g, SQL Server 2000, Sybase 12.x, TERADATA V2 R5, DB2 UDB 7.1,7.2,8.1 EE/EEE, IBM Informix 9.2);
- Ontwerp-tools (ERwin, Oracle Designer, Sybase PowerDesigner, IBM Rational ER, Microsoft VISIO Database, Embarcadero ERStudio);
- ETL tools (PowerMart en PowerCenter);
- Reporting tools (PowerAnalyzer, BusinessObjects, Cognos Impromptu en ReportNet), MicroStrategy en DB2 CubeViews.

Naast deze standaard Xconnects kan een gebruiker een 'eigen' XConnect ontwikkelen voor de ontsluiting van een nog niet ondersteund tool.

PowerAnalyzer

Voor de reporting- en analyse-mogelijkheden van SuperGlue wordt gebruik gemaakt van een speciale versie van Informatica's



Afbeelding 3 : Voorbeeld van een data lineage-rapport in SuperGlue.

eigen PowerAnalyzer. PowerAnalyzer is een op dashboards gebaseerd reporting- en analyse-tool. De versie die binnen SuperGlue gebruikt wordt is uitgebreid met een module om data lineage-rapporten te genereren. Hiermee kunnen ontwikkelaars en beheerders snel een overzicht krijgen van waar een bepaald gegeven binnen alle tools gebruikt wordt. In afbeelding 3 toont een data lineage-rapport.

Naast data lineage-rapporten is het uiteraard ook mogelijk standaardoverzichten van de diverse tool-metadata uit te draaien. Dit is geweldig voor documentatiedoeleinden, want de rapporten kunnen eenvoudig geëxporteerd worden naar HTML en/of PDF. Ook de reguliere zoekfuncties waarmee eenvoudig door de metadata heen gezocht kan worden naar vrije zoektermen zitten natuurlijk in SuperGlue. Een functie die zeker niet onvermeld mag blijven is het *where used rapport*. Een wijzigingsbeheerder kan een impact-analyse uitvoeren door middel van zo'n where used rapport, waarin alle metadata-objecten worden getoond waarin een bepaalde zoekterm voorkomt.
De wijzigingsbeheerder kan zo snel kijken waar in alle systemen bijvoorbeeld de klantcode gebruikt wordt, wanneer de definitie daarvan op het punt staat gewijzigd te worden. Door een where used rapport uit te draaien ziet de wijzigingsbeheerder meteen in welke datawarehouse-dimensies en facts de klantcode gebruikt wordt, door welke ETL-processen ze gebruikt wordt en op welke rapporten ze voorkomen, zie afbeelding 4.

SuperGlue Where-Used Analysis - Property Of Informatica Corporation

Where-used trail objects

Object name	Class	Repository
SALES_VALUE	OracleColumn	Mersche Motors Oracle database
SALES_VALUE	Column	Mersche Motors Business Object
Sum of Sales Value	MeasureObject	Mersche Motors Business Object

Domain: Database Management

Repository: Mersche Motors Oracle database

Class	Object name	Location
OracleColumn	SALES_VALUE	Mersche Motors Oracle database > MM_DATAMART > DM_MM_SALES_
OracleDatatype	NUMBER	Mersche Motors Oracle database > NUMBER
OracleTable	DM_MM_SALES_INVOICES	Mersche Motors Oracle database > MM_DATAMART > DM_MM_SALES_

Domain: Data Integration

Repository: Mersche Motors PowerCenter

Class	Object name	Location
TargetDefinitionPort	SALES_VALUE	Mersche Motors PowerCenter > MM_DM > DM_MM_SALES_INVOICES > S

Domain: Business Intelligence

Repository: Mersche Motors Business Objects

Class	Object name	Location
Class	MM_DATA_DM Measures	Mersche Motors Business Objects > Universe > MM_DATAMART >
Column	SALES_VALUE	Mersche Motors Business Objects > Universe > MM_DATAMART >
DataProviderColumn	Sum of Sales Value	Mersche Motors Business Objects > Document > Monthly Revenue MMDATADM > Dataset for Query 1 with MMDATADM > Sum of Sale
DataType	Number	Mersche Motors Business Objects > Number
DimensionObject	Sales Value	Mersche Motors Business Objects > Universe > MM_DATAMART > Value Mersche Motors Business Objects > Universe > MM_DATAMART > Sales Value
MeasureObject	Accessories Revenue	Mersche Motors Business Objects > Universe > MM_DATAMART > Mersche Motors Business Objects > Universe > MM_DATAMART >
	Sum of Sales Value	Mersche Motors Business Objects > Universe > MM_DATAMART > Mersche Motors Business Objects > Universe > MM_DATAMART > of Sales Value
	Total Revenue	Mersche Motors Business Objects > Universe > MM_DATAMART > Mersche Motors Business Objects > Universe > MM_DATAMART >
	Vehicles Revenue	Mersche Motors Business Objects > Universe > MM_DATAMART > Mersche Motors Business Objects > Universe > MM_DATAMART >
MeasureTable	Measures from DM_MM_SALES_INVOICES	Mersche Motors Business Objects > Universe > MM_DATAMART >
Table	DM_MM_SALES_INVOICES	Mersche Motors Business Objects > Universe > MM_DATAMART >

Repository: Mersche Motors Cognos Impromptu

Class	Object name	Location
ImpromptuSchemaElement	SALES_VALUE	Mersche Motors Cognos Impromptu > mm_cognos > DM_MM_SALES_

Afbeelding 4: Voorbeeld van een where used rapport.

Zo kan gericht en snel bepaald worden hoeveel impact en dus tijd en geld een voorgenomen wijziging gaat kosten. Dit biedt uiteraard enorme voordelen voor een beheerorganisatie, want hoe vaak is het niet voorgekomen dat een rapport niet meer valide was of een ETL-proces vastliep doordat een bron-database tabel gewijzigd was.

Een top-5 van langstlopende ETL-processen kan worden opgesteld

Ook kunnen operationele metadata opgevraagd en gecombineerd worden. Zo kunnen bijvoorbeeld de 'duurste' gegevenselementen bepaald worden op basis van de verwerkingstijd om van bron tot datawarehouse te komen. Ook kan bijvoorbeeld een top-5 van langstlopende ETL-processen worden opgesteld.

Conclusies

Metadata krijgen een belangrijker rol binnen IT-afdelingen, omdat bedrijven meer informatie beschikbaar stellen aan gebruikers en deze informatie uit steeds meer verschillende bronsystemen komt. De daarvoor benodigde integratie-tools bevatten groeiend aantal metadata waardoor het bos voor de ontwikkelaar en beheerder steeds dichter wordt. Het wordt lastiger precies te bepalen wat de impact is van een tabel-wijziging in een database of om te achterhalen waar een bepaald rapport-gegeven uiteindelijk vandaan komt. Het gebruik van een metadata management tool kan op dat gebied veel verlichting bieden voor zowel gebruikers als ontwikkelaars en beheerders. Informatica's SuperGlue is daarbij slechts één van de op dit moment op de markt zijnde tools.

Rick Mutsaers (rick.mutsaers@ordina.nl) is senior BI Consultant bij Ordina VisionWorks.